Change-Point Detection for High-Dimensional Data

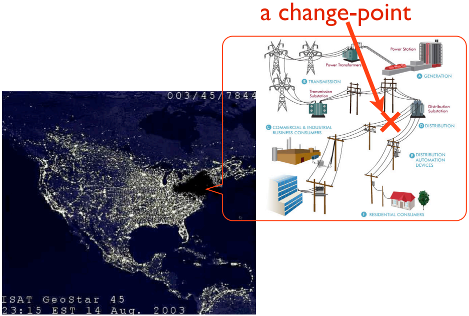
Yao Xie* and Rebecca Willett†

* Georgia Institute of Technology, † University of Wisconsin, Madison

Asilomar Conference on Signals, Systems, and Computers 2013

November 5, 2013

Power network monitoring



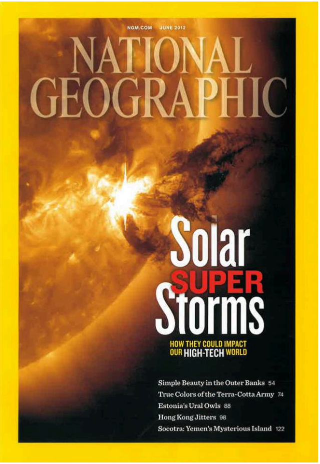
Northeast blackout, 2003

- Northeast blackout of 2003 is a widespread power outage in the US
- caused by

  "...did not recognize the deteriorating condition of the system [in the Ohio part]."

- **change-point**:
  break-down of power line
- detect the change-point in real-time might prevent large-scale blackout

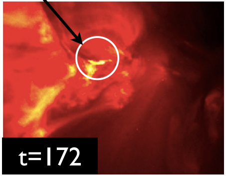
Solar flare detection



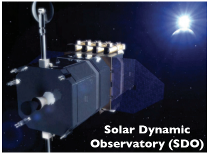
June, 2012.

- solar storm: a large explosion in the sun's atmosphere
- a direct hit by solar storm can shut down power lines and destroy airplane communications
- predict major solar storms: detecting small **solar flares** (sudden brightening over sun's surface)
- **change-point**: solar flare

a change-point



Data for solar flare detection



source: NASA SDO

- images taken by helioseismic sensors mounted on the satellite

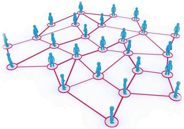
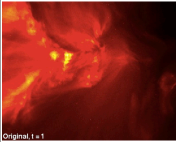
Detecting solar flares is hard

- **high-dimensional**: each frame has $232 \times 292 = 67744$ pixels
- **large volume**: 130Mbps, acquire $\sim$ 11 terabytes/day
- **dynamic**: sun surface (background) changing
- **quick detection** of a small solar flare
- **low complexity**: real-time detection necessitates computationally efficient algorithms
- **missing data**: sensors may be overloaded
- **communication and storage**

marinecorpstimes.com

High-dimensional streaming data

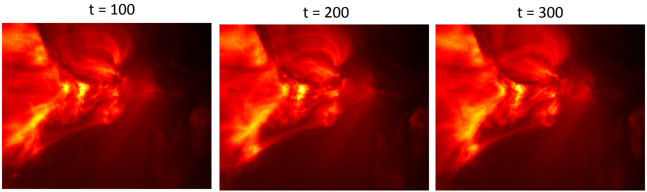
- Each dimension is sequence of data measured by a "**sensor**"
- Large number of "**sensors**"

applications	EEG brain imaging 	power network monitoring 	social network 	solar flare detection 
sensors	EEG sensors 	smart meters 	"sensor": each person	"sensor": each pixel
# sensors	$\sim 21 - 128$	$\sim 10,000$ in a subnet	thousands to millions	$\sim 60,000$ pixels

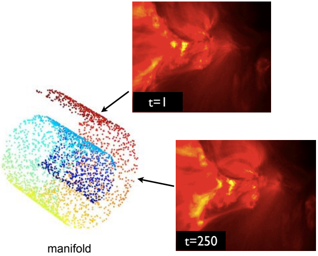
kk.org, wikipedia, ubiu.co.kr, sensysnetworks.com, P. Varaiya

Background is dynamic and low-dimensional

Background is dynamic...



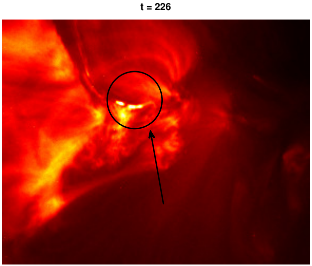
... and low-dimensional

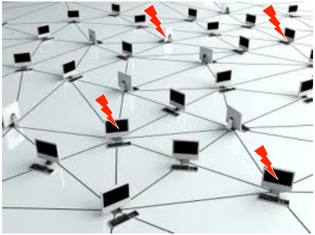


Tenenbaum, Silva and Langford 2000, Roweis and Saul 2001.

implicitly reduced Picard iteration for online manifold learning: Peng and Mosheni, 2012.

Anomaly signal is sparse





change-point only affects a small subset of sensors

today.mccombs.utexas.edu

Detect sparse changepoint in low-dimensional signal

After anomaly occurs:

signal = background + anomaly

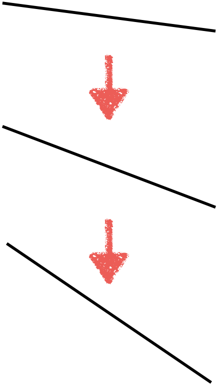
low-dimensional slowly time-varying	sparse appear abruptly in time

Goal: Detect occurrence of a sparse signal in a slowly-time varying low-dimensional background **in real time**

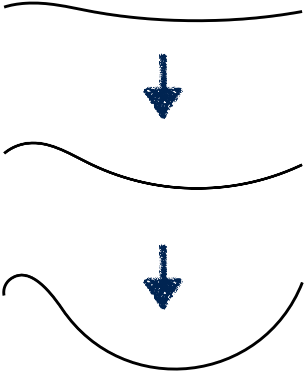
Low-dimensional and slowly time-varying signal

many high-dimensional signals
exhibit low-dimensional structures

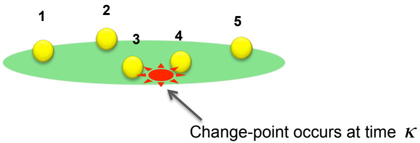
rotating hyperplane



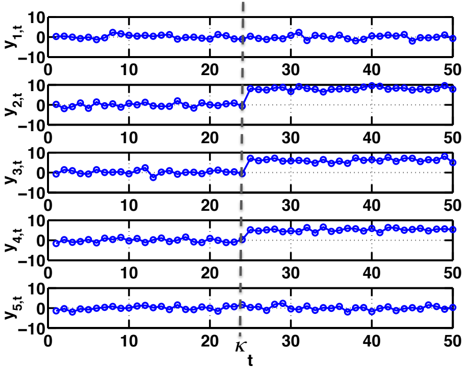
time-varying manifold



Model



1
2
3
4
5
Change-point occurs at time κ



$y_{1,t}$
$y_{2,t}$
$y_{3,t}$
$y_{4,t}$
$y_{5,t}$
10
0
−10
0
10
20
30
40
50
κ
t

Formulation

Data $\{y_{n,t}\}_{n=1,\cdots,N, \quad t=1,2,\cdots,}$

N: size of the dimension

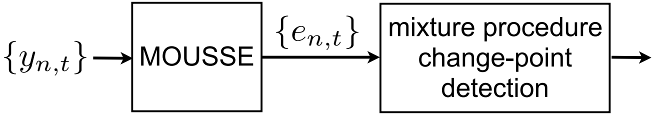
$$\mathrm{H}_0: \quad y_{n,t} = x_{n,t} + w_{n,t}, \quad n = 1, \cdots, N, \quad t = 1, 2, \cdots \tag{1}$$

$$\mathrm{H}_1: \quad \begin{cases} y_{n,t} = x_{n,t} + w_{n,t}, & n \in \mathcal{S}, \quad t = 1, \cdots, k; \\ y_{n,t} = x_{n,t} + \mu_n + w_{n,t}, & n \in \mathcal{S}^c, \quad t = k+1, \cdots \end{cases}$$
$$w_{n,t} \sim \mathcal{N}(0, 1)$$

- Unknown parameters
  - $x_{n,t} \in \mathcal{M}_t$: slowly time-varying, low-dimensional
  - μ_n: changepoint amplitude
  - $\mathcal{S}$: subset of sensors affected
  - k: time when changepoint occurs

Two-stage algorithm

- **Track** time-varying component $x_{n,t}$ using $y_{n,t}$
- **Detect** changepoint as soon as possible after it occurs



Tracking dynamic manifolds: MOUSSE algorithm [2]

- simple online manifold learning algorithm:

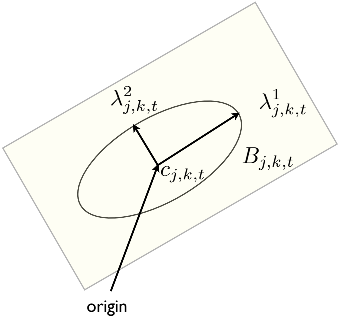
  (a) learn low-dimensional approximation sequentially

  (b) deal with missing data

- union-of-subsets: $\widehat{\mathcal{M}}_t = \bigcup_{(j,k)} \mathcal{S}_{j,k}$,

- subset: ellipsoid on hyperplane

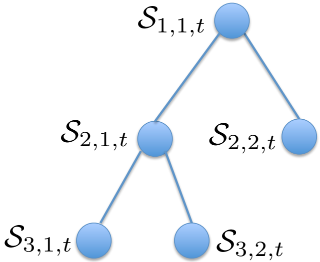
$$\mathcal{S}_{j,k,t} = \{v : v = \overbrace{B_{j,k,t}}^{\text{subspace basis}} z + \overbrace{c_{j,k,t}}^{\text{offset}}$$
$$z^\top \underbrace{\Lambda_{j,k,t}^{-1}}_{\text{spread}} z \leq 1$$
$$z \in \mathbb{R}^d\}$$

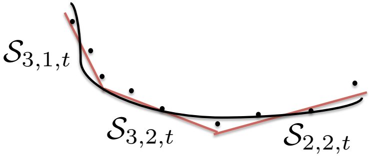


Xie, Huang and Willett, Change-point detection for high-dimensional time series with missing data, IEEE Journal Sel. Topics Signal Processing, Feb. 2013

Online learning in MOUSSE

- subsets are multiscale and organized in tree structure





- online-learning

$$\cdots \xrightarrow{y_{n,t}} \widehat{\mathcal{M}}_t \xrightarrow{y_{n,t+1}} \widehat{\mathcal{M}}_{t+1} \xrightarrow{y_{n,t+2}} \cdots$$

- update subset parameters and tree structure

Subspace tracking: Balzano et.al. 2011, Chi et.al. 2012
Geometric wavelets: Allard, Chen and Maggioni, 2011

Robust to missing data

(Xie, 2013) Given union-of-subset: $\bigcup_{i=1}^{K} \mathcal{S}_i$. If entries "missing at random",

1) enough number of observed entries:

$$|\Omega| \geq \max_{i=1}^{K} \left\{ \frac{8}{3} d_i \mathrm{coh}(U_i) \log(2 d_i (K-1)/\delta) \right\},$$

2) enough "distance margin" between "optimal" subset and other subsets

$$c_\star \sin(\theta_\star) + a_\star \rho(x, \mathcal{S}_\star) < c_i \sin(\theta_i) + a_i \rho(x, \mathcal{S}_i), \quad \forall i \neq \star,$$

then with prob. $1 - 6\delta$, projection of data can be well-approximated:

$$\| U_\star^\dagger (x - c_\star) - U_\Omega^\dagger (x_\Omega - c_\Omega) \|^2 \leq (1+\eta)^2 \underbrace{\frac{D - |\Omega|}{(1-\gamma)^2}}_{\text{undersampling}} \underbrace{\frac{d_\star \mathrm{coh}(U_\star)}{D}}_{\text{subspace property}} \underbrace{\|q^\star\|^2}_{\text{best approx. err.}}$$

Coherence of basis: $\mathrm{coh}(U_i) = (D/d_i) \max_j \|U_i U_i^\dagger e_j\|^2$

Proof ideas from "matrix completion", Candes and Tao, 2009, "subspace matching" Balzano, Recht and Nowak 2011.

MOUSSE residuals

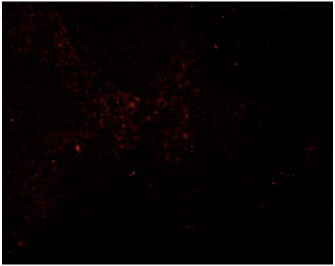
- estimate time varying component

$$\hat{x}_{n,t} = \text{projection of } y_{n,t} \text{ onto } \widehat{\mathcal{M}}_{t-1}$$

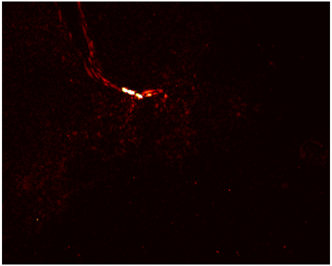
- residuals

$$e_{n,t} = y_{n,t} - \hat{x}_{n,t},$$

before change-point,
residual small



after change-point,
around solar flare, residual large



Multi-sensor changepoint detection with sparsity

If tracking algorithm is working well, model for the residual

$$\mathrm{H}_0: \quad e_{n,t} = w_{n,t}, \quad n = 1, \cdots, N, \quad t = 1, 2, \cdots$$

$$\mathrm{H}_1: \quad \begin{cases} e_{n,t} = w_{n,t}, & n \in \mathcal{S}, \quad t = 1, \cdots, k; \\ e_{n,t} = \mu_n + w_{n,t}, & n \in \mathcal{S}^c, \quad t = k+1, \cdots \end{cases}$$

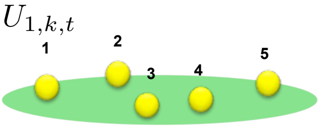
$$w_{n,t} \sim \mathcal{N}(0, 1)$$

- only few entries of $\{\mu_n\}$ are non-zero

Exact likelihood procedure

- each sensor forms GLR statistic

$$U_{n,k,t} = \frac{(\sum_{i=k+1}^{t} y_{n,i})^2}{t-k}$$



- sum over set of affected sensors

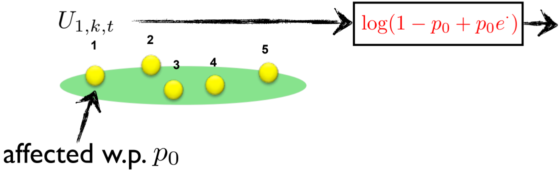
$$T = \inf\{t : \max_{k<t} \max_{\mathcal{S}\in\Omega} \sum_{n\in\mathcal{S}} U_{n,k,t} \geq b\}$$

- $\mathcal{S}$ unknown, have to search all possible subsets of $\{1, \cdots, N\}$
- number of possible subsets $|\Omega| = 2^N$, exponential in N

Mixture procedure

- exploit signal **sparsity**:
  introduce **mixture model**
  - assume each sensor affected with probability p_0
  - p_0 is a guess for fraction of affected sensors ($p = |\mathcal{S}|/N$)
- mixture procedure

$$T_{\text{mix}} = \inf\{t : \max_{k<t} \sum_{n=1}^{N} \log(1 - p_0 + p_0 e^{U_{n,k,t}^2/2}) \geq b\}$$



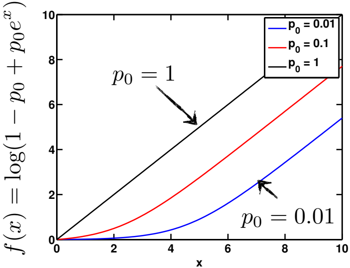
Xie and Siegmund, Sequential multi-sensor change-point detection, Annals of Statistics, 2013.

Soft-thresholding

- select sensors with useful information by

  "soft-thresholding" the GLR statistic (not the signal)

  soft-threshold function: $f(x) = \log(1 - p_0 + p_0 e^x)$



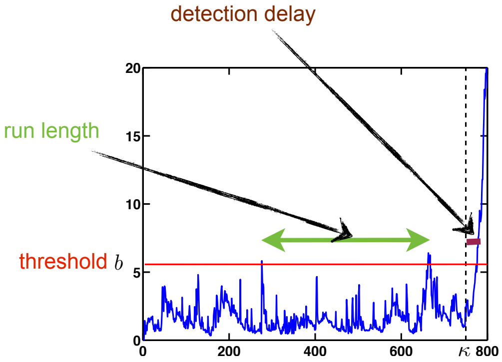
GLR statistics

unaffected sensor: $U_{n,k,t} \approx 1$

affected sensor: $U_{n,k,t} \approx (t-k)\mu_n^2$

$$T_{\text{mix}} = \inf\{t : \max_{k<t} \sum_{n=1}^{N} \log(1 - p_0 + p_0 e^{U_{n,k,t}^2/2}) \geq b\}$$

Performance metrics



- average run length (ARL):

$$\mathbb{E}^\infty\{T\}$$

- expected detection delay (EDD):

$$\sup_k \operatorname{ess\,sup} \mathbb{E}^k\{T - k | T > k\}$$

Average Run Length (ARL)

Theorem (Xie and Siegmund, 2013)

As $b \to \infty$, $N \to \infty$, with b/N fixed, for $\theta : \dot{\psi}(\theta) = b/N$,

$$\text{(ARL)} \quad E^{\infty}\{T\} = \frac{\theta[2\pi\ddot{\psi}(\theta)]^{1/2}}{c(N)\gamma(\theta)N^{1/2}} \exp\{N[\theta\dot{\psi}(\theta) - \psi(\theta)]\} + o(1).$$

b: threshold, N: number of sensors.

- complicated expression, but approximation highly accurate
- given ARL, b can be efficiently determined
- $\text{ARL} = \mathcal{O}(e^{b-N/2})$

Expected Detection Delay (EDD)

Theorem (Xie and Siegmund, 2013)

As $b \to \infty$,

$$(\text{EDD}) \quad \mathbb{E}^0\{T\} = (\Delta^2/2)^{-1}[b + 2\rho(\Delta) - |\mathcal{S}| \log p_0 - |\mathcal{S}|/2 - (N - |\mathcal{S}|)\mathbb{E}\{g(U, p_0)\} - 1 - \Delta^2/4 + o(1)].$$

- signal energy:

$$\Delta^2 = \sum_{n \in \mathcal{S}} \mu_n^2$$

  μ_n, $n \in \mathcal{S}$: post-change mean for affected sensors

- $\Delta^2/2 =$ **Kullback-Leibler (K-L) divergence**
- EDD $= \mathcal{O}(\frac{b}{\Delta^2/2})$

Sharp approximations

Table : Average run length (ARL) of T_{mix}

p_0	b	Theory	Monte Carlo	% difference
0.1	19.5	5000	4968	0.6%
0.1	20.4	10001	10093	0.9%
0.03	12.7	5001	4830	3.4%
0.03	13.5	10001	9948	0.5%

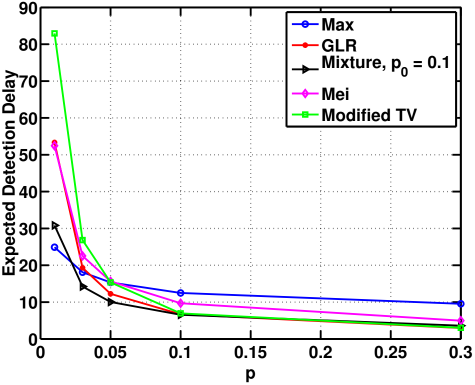
Table : Expected detection delay (EDD) of T_{mix}, ARL $\approx$ 5000, $\mu_n = 1$.

p	p_0	Theory	Monte Carlo	% difference
0.1	0.1	7.2	6.7	6.9%
0.03	0.1	13.9	14.3	2.9%
0.03	0.03	13.9	14.2	2.1%

mixture procedure is first-order asymptotic optimal

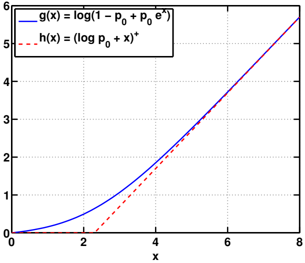
Detection delay

- Performance of mixture procedure is better



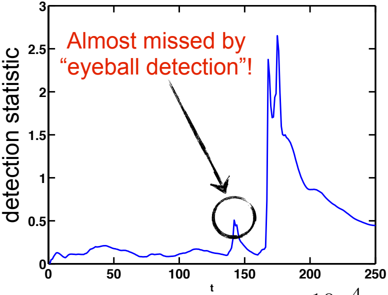
Hard-thresholding

- Avoid numerical stability issue in $\log(1 - p_0 + p_0 e^{x^2/2})$
- We have analytic approximate ARL of hard-thresholding mixture procedure.

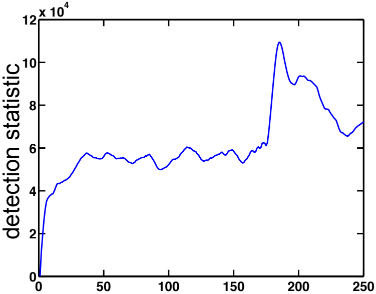


$$T_{mix,h} = \inf\left\{ t : \max_{0 \le k < t} \sum_{n=1}^{N} [U_{n,k,t} + \log(p_0)]^+ \ge b \right\}$$

Apply on solar data



mixture procedure, $p_0 = 10^{-4}$



without exploiting sparsity

Future work

- Detect in real-time
  - abrupt occurrence of sparse changepoint (anomaly)
  - in slowly time-varying low-dimensional background
- combines
  - tracking of low-dimensional background
  - sparse changepoint detection
- when there is correlation among the sensors affected by the changepoint, how do we capture that in the detection statistic